Estimands, Missing Data, and Sensitivity Analysis

Geert Molenberghs

geert.molenberghs@uhasselt.be & geert.molenberghs@kuleuven.be

Interuniversity Institute for Biostatistics and statistical Bioinformatics (I-BioStat)

Universiteit Hasselt & KU Leuven, Belgium

www.ibiostat.be

U Penn Symposium

April 18, 2018

Survey Sampling 101

Survey population: The collection of units (individuals) about which the researcher wants to make quantitative statements.

Sample frame: The set of units (individuals) that has non-zero probability of being selected.

Sample: The subset of units that have been selected.

Probability sampling: The family of probabilistic (stochastic) methods by which a subset of the units from the sample frame is selected.

Design properties: The entire collection of methodological aspects that leads to the selection of a sample.

Sample size: The number of units in the sample.

Analysis and inference: The collection of statistical techniques by which population estimands are estimated.

Examples: estimation of means, averages, totals, linear regression, ANOVA, logistic regression, loglinear models.

Estimand: The true population quantity (e.g., the average body mass index of the US population).

Estimator: A (stochastic) function of the sample data, with the aim to "come close" to the estimand.

Estimate: A particular realization of the estimator, for the particular sample taken (e.g., 22.37).

Your M.o.t.R. Clinical Trial

- Setting:

Potential outcomes	(T_{0j}, T_{1j})
Individual treatment effect	$\Delta_{Tj} = T_{1j} - T_{0j}$
Expected treatment effect	$\beta = E(T_{1j} - T_{0j})$

- No missing data $\Longrightarrow$ 50% of missing data
- Fair to say: **Estimand** is $\beta = E(T_1 - T_0)$ in **population**
- **Randomization:** Treatment effect estimable from observed data:
- **Estimator:** $\overline{T_1} - \overline{T_0}$

- **Information coming from:**
  - ▷ data
  - ▷ design
  - ▷ ~~assumptions~~
- Would be different in an **epidemiological** study

Surrogate Endpoints Evaluation: Potential Outcomes

Alonso, Van der Elst, Molenberghs (Statistical Modeling 2016)

- Setting:

Potential outcomes	(T_{0j}, T_{1j})
Individual causal effect	$\Delta_{Tj} = T_{1j} - T_{0j}$
Expected causal effect	$\beta = E(T_{1j} - T_{0j})$
Surrogate	S_j

- **Predictive causal association:**

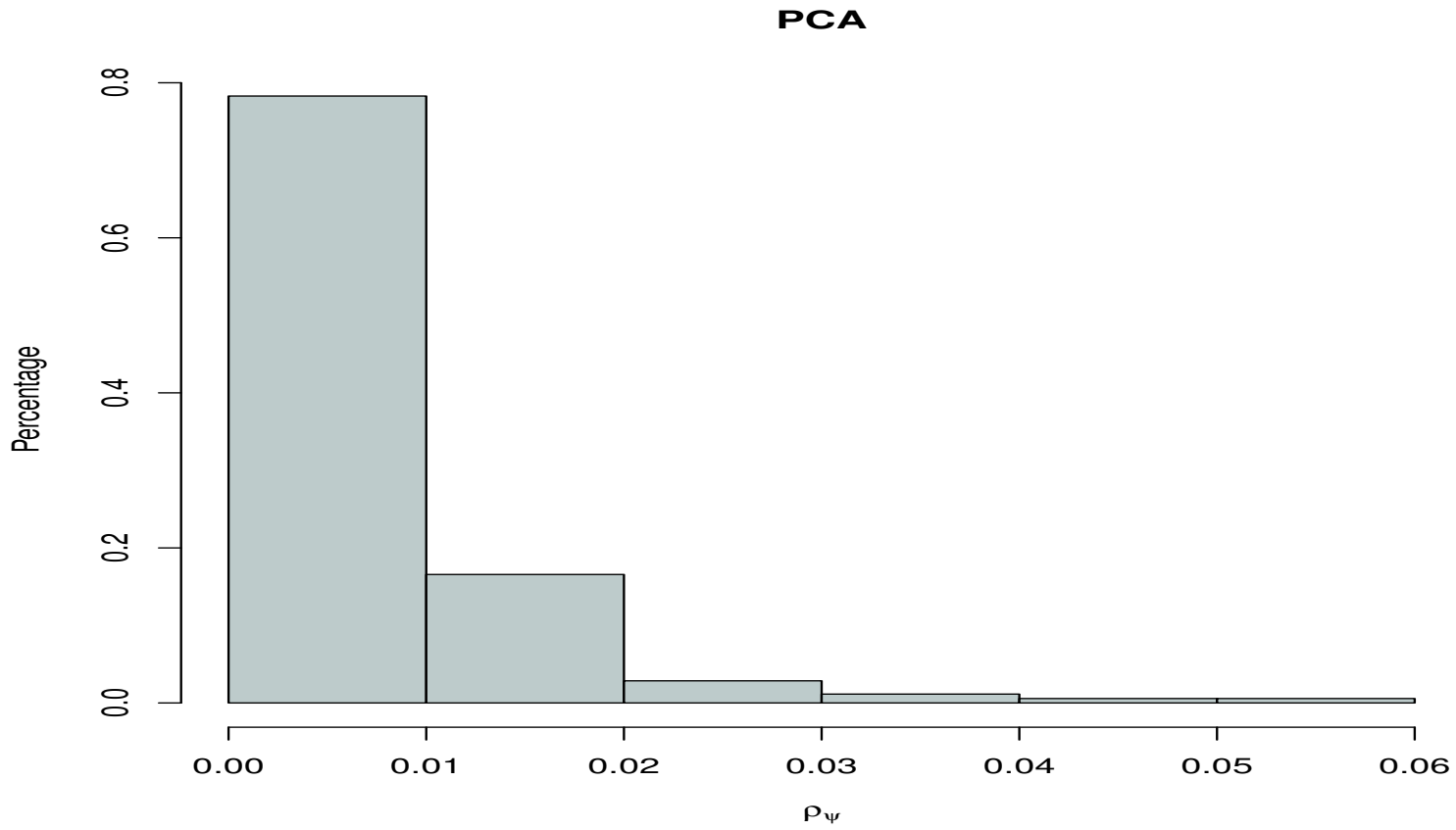
$$\rho_\psi = \mathsf{corr}(\Delta_{Tj}, S_j)$$

- **(Un)identifiability**:

$\rho_{T_0T_1}$ **not identifiable**

- **Information coming from:**
  - ▷ data
  - ▷ design
  - ▷ assumptions $\longrightarrow$ **sensitivity**

⇒ **Sensitivity analysis** for age-related macular degeneration trial:



Surrogate Endpoints Evaluation: Full Causal Paradigm

Alonso et al. *(Biometrics 2015)*

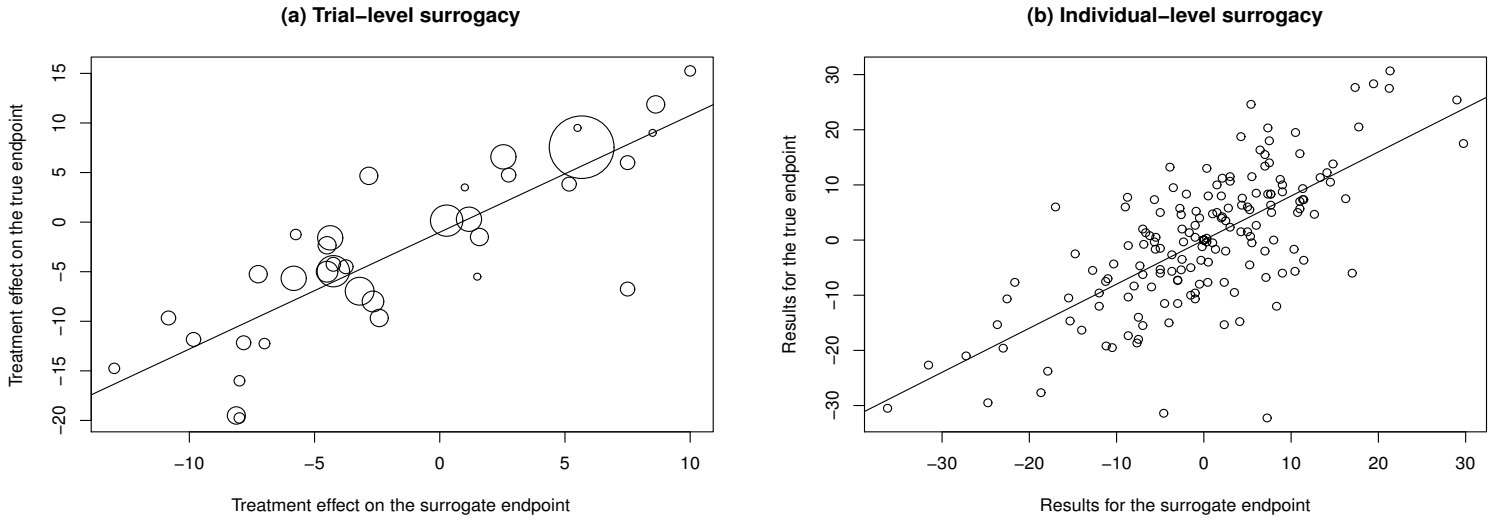
- Setting:

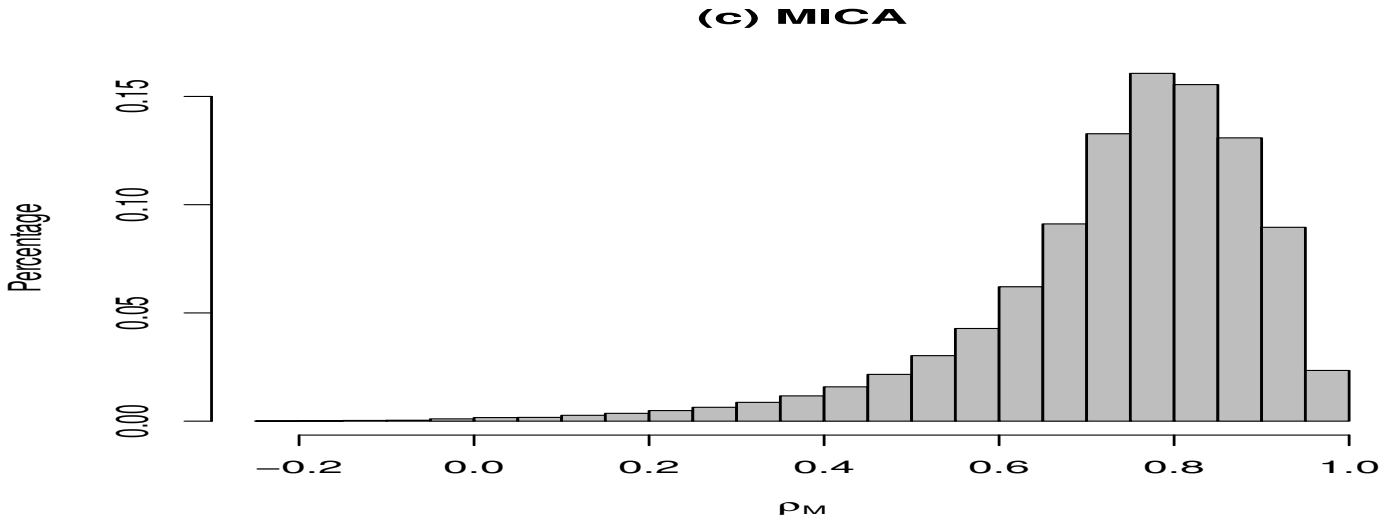
Treatment potential outcomes	(T_{0j}, T_{1j})
Treatment individual causal effect	$\Delta_{Tj} = T_{1j} - T_{0j}$
Surrogate potential outcomes	(S_{0j}, S_{1j})
Surrogate individual causal effect	$\Delta_{Sj} = S_{1j} - S_{0j}$

- **Individual causal association (ICA):**

$$\rho_{\Delta} = \mathsf{corr}(\Delta_{Tj}, \Delta_{Sj})$$

- **Joint distribution unidentifiable**
- Capture assumptions in **causal diagrams** $\longrightarrow$ reduced forms of ρ_{Δ}
- **Information coming from:**
  - ▷ data
  - ▷ design
  - ▷ assumptions $\longrightarrow$ **sensitivity**
- Meta-analytic version in multiple trials


(a) Trial−level surrogacy
Treatment effect on the true endpoint
Treatment effect on the surrogate endpoint
−20
−15
−10
−5
0
5
10
15
−10
−5
0
5
10
(b) Individual−level surrogacy
Results for the true endpoint
Results for the surrogate endpoint
−30
−20
−10
0
10
20
30
−30
−20
−10
0
10
20
30


(c) MICA
Percentage
0.00
0.05
0.10
0.15
−0.2
0.0
0.2
0.4
0.6
0.8
1.0
ρM

Terms of Enrichment

Enriched data	
Coarse data	**Augmented data**
Incomplete data	**Randomized studies**
Censored data	Random effects
Joint models	Latent classes
Grouped data	Latent variables
Non-compliance	Mixtures

Increasing Complexity

- **Standard clinical trial:** design compensates for what is unobserved
- **Surrogacy:** augmentation: sensitivity is **design**-based
- **Incomplete data/non-compliance:** coarsening: sensitivity is **(non-)observation**-based
  - ▷ (Subjective) choices unavoidable
  - ▷ Interference of intercurrent events
  - ▷ Scenarios needed about $f(\boldsymbol{y}_i^m|\boldsymbol{y}_i^o, \boldsymbol{x}_i, \boldsymbol{\theta})$ (Devan, p. 9)
  - ▷ **Such scenarios should preserve estimand**
  - ▷ Easy and elegant with MI

Concluding Reflections

- Devan starts with the right point question: **WHY?**
- Both: taxonomy is a **GOOD** thing
  - ▷ Devan: Proper definitions needed: objective/question — endpoint — estimand
  - ▷ Tom: principal stratification *can* be of help
- **Sensitivity analysis** ⊚ **Estimands**